ChinaXiv合作期刊

chinaXiv:201903.00229v1

基于样条函数的恒星光谱自动归一化方法

罗锋[1,2]，刘超[1,2]，赵永恒[1,2]

(1, 中国科学院国家天文台，北京100101；2, 中国科学院大学，北京100101)

摘要：恒星的观测谱一般由连续谱、谱线和噪声组成。其中连续谱是由黑体辐射导致的辐射流量随波长变化的光滑连续光谱。光谱分类以及恒星物理参数估计等研究都依赖于连续谱及谱线信息的准确提取。因此光谱数据处理的主要工作就是拟合连续谱，并通过对光谱进行归一化来提取谱线特征。目前连续谱拟合的方法主要有多项式拟合、中值滤波、小波滤波等。已有的方法在低信噪比，宇宙线信号干扰，存在发射线等情况下，都有不同程度的局限性，主要体现在鲁棒性和准确度上。目前，对于LAMOST的10^7条光谱没有自动化方法应用到归一化上，在天文数据雪崩的时期，研究并开发一种能够适用于更广的温度、信噪比及波长覆盖范围的，具有更好普适性并能够自动化处理的恒星光谱归一化算法，显得十分迫切。在仔细分析不同类型光谱的基础上，提出了一种基于固定窗口划分的连续谱拟合方法。该方法对光谱中能够体现连续谱特征的数据点进行筛选提取，通过细微地控制样条函数平滑度来产生更加准确的连续谱。使用LAMOST中不同光谱型、温度范围、波长覆盖范围的光谱进行实验，结果表明文章提出的算法具有良好的精度和普适性。

关键词：连续谱归一化；LAMOST；恒星光谱；样条函数

中图分类号：　　文献标识码：　　文章编号：

0　引言

作为国家重大科学工程，LAMOST 突破了天文望远镜大视场与大口径难以兼得的难题[1]，成为目前国际上口径最大的大视场望远镜，是我国光学望远镜研制的又一里程碑。截至 2018 年 6 月 LAMOST 完成了七年的巡天观测，获取光谱数首次超千万量级，也成为世界上第一个获取光谱数超千万的巡天项目。**遗憾的是，对于 LAMOST 的 10^7 条光谱目前没有自动化方法应用到归一化上，本文主要对这一步骤进行研究。**

每条观测谱都由连续谱、谱线和噪声组成[2]。通过谱线可以对恒星的化学组成进行分析，并且由于多普勒效应，谱线还携带着视向速度的信息。因此，从原始光谱中准确的提取谱线是恒星光谱研究及光谱数据处理工作中的重要步骤，对后续的研究有着重要意义。

要消除连续谱首先要对连续谱进行拟合。已有的连续谱拟合方法主要有：多项式拟合、中值滤波、小波滤波、形态滤波器等[3]。也有利用天文软件进行半自动处理的方法：**首先人工筛选提取认为合适的数据点**[4]**并选择函数形式**，然后由软件进行多项式拟合，最后得到连续谱。这种方法依赖人工参与，无法满足目前巡天项目中海量光谱的实时处理需求。国外的斯隆巡天

*基金项目：国家自然科学基金面上项目(11873057)资助.
收稿日期：修订日期：
作者简介：罗锋，男，博士生，研究方向：天文技术与方法。Email: luofeng@nao.cas.cn

项目（SDSS）中的数据处理程序 SSPP（the segue stellar parameter pipeline）采用的方法是分段多项式拟合[5]。首先将光谱分为红端和蓝端两个部分，去除强巴尔末线系以后，对蓝端使用 9 阶，红端使用 4 阶多项式分别拟合，丢弃 3 个标准差以外的数据点。然后拼接两段连续谱再进行一次 9 阶多项式拟合，得到最终的连续谱。这种方法稍显繁琐，且计算量大。

本文提出的方法将光谱划分为数个固定的（包含相同像素数量）窗口，窗口内选中值点，然后通过筛选策略丢弃部分参考点，能够非常有效地避开发射线、吸收线及宇宙线。并且通过对参考点之间“夹角”的控制，可以使用非常小的平滑度来构造样条曲线，实现对连续谱的精确拟合。

1　方法介绍

本文提出的连续谱归一化方法大体上包含三个步骤：首先，将一条待处理光谱根据数据点总数划分为数个窗口，每个窗口选出流量中值点形成参考点集；其次，制定了一些筛选策略，对参考点集中有可能影响拟合结果的数据点进行丢弃。最后，根据参考点流量最大值所在波长位置选取对应的样条函数平滑度经验值进行拟合，得到连续谱。

1.1　划分窗口构造参考点集

为了有足够的数据点作为拟合函数的控制点，根据输入的待处理光谱数据点总数，采用如下策略划分窗口。设数据点总数为 n，每个窗口包含像素数，即窗宽为 w：

表 1　窗宽与数据点总数之间的对应关系

Tab 1. The relationship between windows width and the total number of data points

n	0~1000	1000~1500	1500~2500	2500~3500	3500~4500	4500~ n
w	13	15	17	19	21	23

根据上表确定窗宽以后对光谱进行窗口划分，然后遍历所有窗口，每个窗口内选取流量值等于窗口中所有点中值的像素作为参考点，并将其加入参考点集 S_{rp}。

1.2　筛选参考点

由上一步骤构造的参考点集 S_{rp} 只是体现了光谱的大致形状，其中有些点可能位于吸收线或者发射线上，如图：

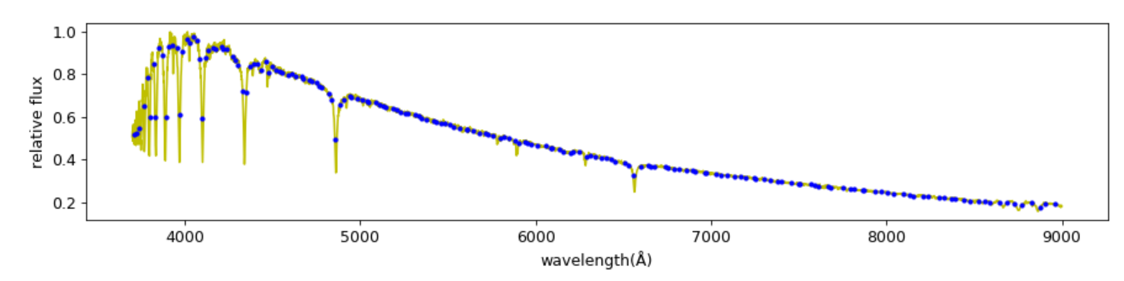


图 1　有些参考点位于吸收线上

Fig. 1 Some reference points are on the absorption line

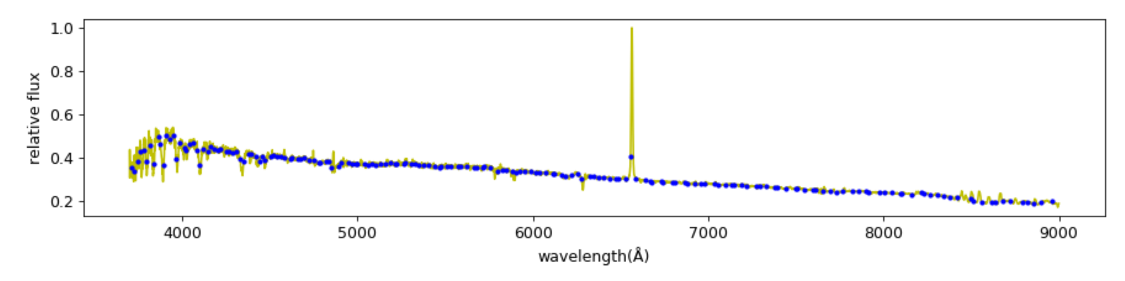


图 2　有些参考点位于发射线上

Fig. 2 Some reference points are on the emission line

chinaXiv:201903.00229v1

chinaXiv:201903.00229v1

连续谱理论上应非常接近黑体辐射谱，而黑体谱可以划分为两个单调区间，峰值左边是单调增区间，右边是单调减区间[6]，所以我们希望用来构造拟合曲线的参考点能够符合并体现这一特征。为了实现这一目标，采取了两个步骤对参考点集再次筛选。

(1).丢弃局部极小值点：

设 S_{rp} 中一个参考点的索引为 i，流量为 $f(i)$，若 i 满足：

$$f(i-1) > f(i)\text{并且}f(i) < f(i+1)\text{，}i >= 1 \tag{1}$$

则将参考点 i 从 S_{rp} 中删去。每次从 i=1 开始到遍历完 S_{rp} 中所有参考点结束为一趟，在一趟丢弃结束的时候，可能还会有些点成为新的局部极小值点，此时需要再丢弃一趟。设丢弃的趟数为 t，我们需要保证以下两点：对连续谱正常的光谱，t 次丢弃之内 S_{rp} 中元素数量已经不再减少；对于连续谱异常[7]的光谱，t 次丢弃以后 S_{rp} 中还有足够的参考点用来构造拟合曲线。实验表明，合适的 t 取值为$t = len(S_{rp})//30 + 2$。

经这一步骤处理后，图 1 和图 2 中两条光谱的参考点集如下图：

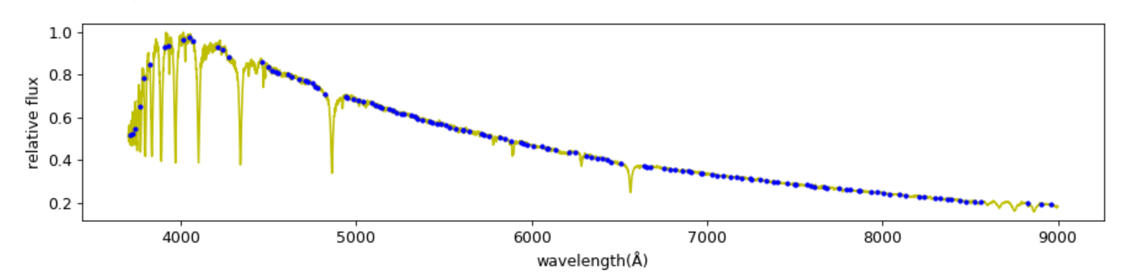


图 3 图 1 中光谱经 t 次丢弃局部极小值以后的参考点集

Fig. 3 The reference point set after the local minimum is discarded t times in the spectrum in Fig. 1

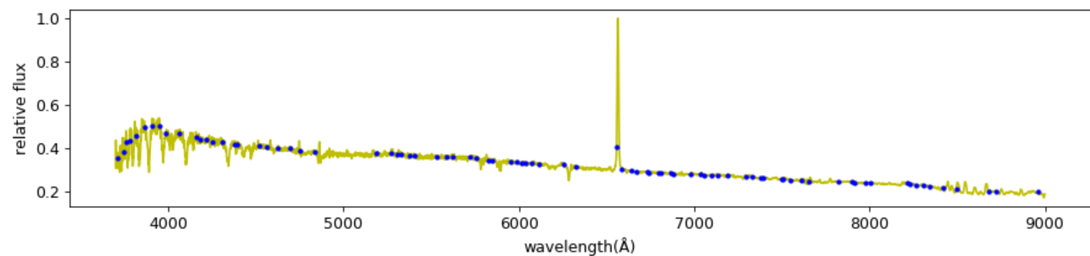


图 4 图 2 中光谱经 t 次丢弃局部极小值以后的参考点集

Fig. 4 The reference point set after the local minimum is discarded t times in the spectrum in Fig. 2

(2).丢弃“夹角”较大的点：

为了获得更加精确的连续谱拟合曲线，我们需要较小的平滑度来构造样条函数。这时，如果有的参考点和其左右两点连线构成的夹角过大，则会给拟合曲线带来较大的偏离，这样的参考点应从 S_{rp} 中删去。如图 5 所示。

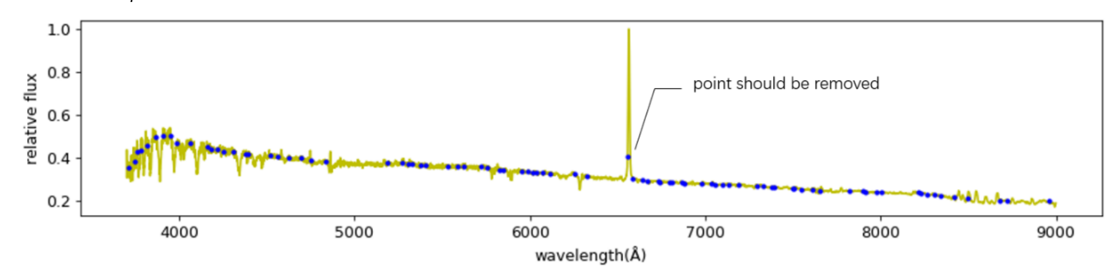


图 5. 位于发射线上的那个数据点应从参考点集中删去

Fig 5 The data point on the emission line should be removed from the reference point set

ChinaXiv合作期刊

chinaXiv:201903.00229v1

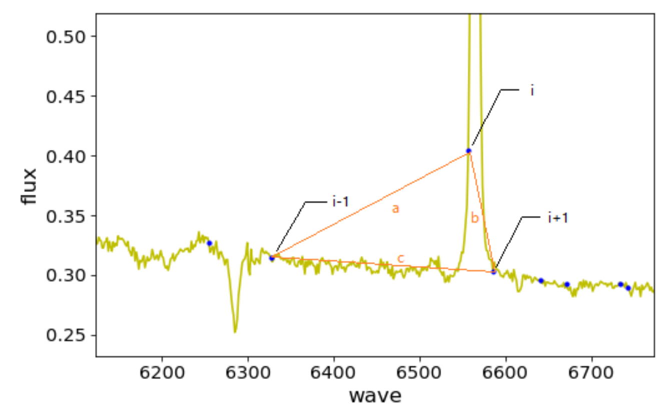


图 6 使用余弦定理计算夹角
Fig. 6 Calculate the angle by use of Cosine theorem

如图 6 所示，在参考点集中取一点 i, 设其对应的波长为$w(i)$，流量为$f(i)$，与它左右两点间连线的夹角为$\theta(i)$，组成的三角形边长分别为a,b,c。在几何上，有：

$$a = \sqrt{[w(i) - w(i-1)]^2 + [f(i) - f(i-1)]^2}, \quad (2)$$

$$b = \sqrt{[w(i+1) - w(i)]^2 + [f(i+1) - f(i)]^2}, \quad (3)$$

$$c = \sqrt{[w(i+1) - w(i-1)]^2 + [f(i+1) - f(i-1)]^2} \quad (4)$$

根据余弦定理便可计算出$\theta(i)$的值。实际上，流量值并不对应长度单位，w 和 f 的单位很难统一。同时，由于我们的光谱数据中流量是未定标的相对流量，直接用余弦定理计算出的夹角会随着光谱流量值所在的区间不同，以及光谱数据处理中必要的乘除运算而变化。为此，我们做如下处理：

对于 S_{rp} 中的一个参考点 i，设其与左边数据点波长数值的差为wd^-，右边为wd^+，则：

$$wd^- = w(i) - w(i-1) \quad (5)$$

$$wd^+ = w(i+1) - w(i) \quad (6)$$

同理，对于流量值[2]：

$$fd^- = f(i) - f(i-1) \quad (7)$$

$$fd^+ = f(i+1) - f(i) \quad (8)$$

引入比例因子r_w, r_f，令：

$$r_w = \frac{wd^-}{wd^- + wd^+} \quad (9)$$

$$r_f = \frac{fd^-}{fd^- + fd^+} \quad (10)$$

最后，计算判据$r = |r_w - r_f|$的值并引入阈值r_0，若$r > r_0$，则认为点 i 属于夹角较大的点，会给精确拟合带来干扰，应从 S_{rp} 中除去。通过实验，我们发现效果较好的r_0取值为 0.5。

如此，通过比值来消去量纲，则不论 w 与 f 取何种单位，r_w, r_f 都是不变的。使用r_w, r_f 作为判断依据来寻找这样的参考点更为可靠。图 5 中光谱经此步骤处理后的结果如图 7 所示。

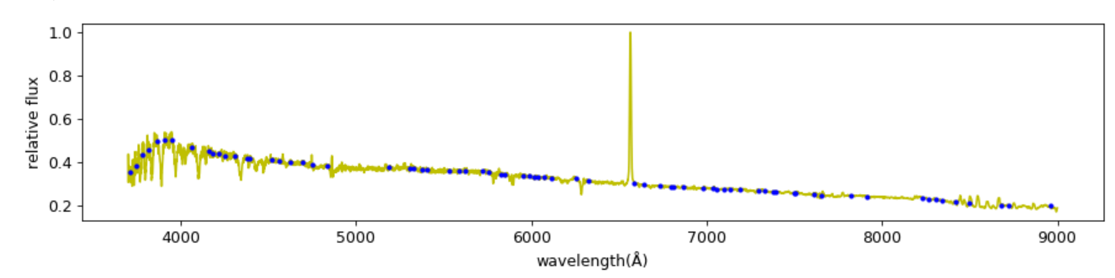


图 7 位于发射线上的参考点被成功移除
Fig.7 The reference point on the emission line was successfully removed

ChinaXiv合作期刊

chinaXiv:201903.00229v1

1.3 平滑度计算

样条函数[8]拟合效果的好坏，除了与参考点的选取有关，还决定于平滑度的选择。对于同一条光谱，同样的参考点集，不同的平滑度会有不同的效果，如图 8 所示。

较小的平滑度可以更好的照顾到光谱图像上起伏的细节，但过小就容易产生振动频繁的现象，使拟合曲线无法体现连续谱；较大的平滑度不会有剧烈的振动，但是过大会使拟合曲线过于平缓，也无法很好的体现连续谱。应该指出的是，对于同一个平滑度，在不同的流量区间，样条曲线也会有不同的表现。我们在拟合之前，都将光谱数据进行了最大值标准化，即 $flux = flux/max(flux)$。

经实验分析，温度越高的光谱需要越小的平滑度来照顾光谱在蓝端较急剧的转折，温度越低的光谱需要越大的平滑度来弱化参考点上下起伏的趋势。而一条光谱的温度在连续谱上体现为流量峰值所在的位置。我们采用在参考点集中选取流量最大值，用其所在波长位置来估计该条光谱的温度区间，选择对应的平滑度s_0。并且，为了能够适用于局部归一化的场合，对于参考点较少的情况，需要更小的平滑度。为此引入默认值为 1 的比例系数 c，用于在参考点较少时将之与s_0相乘的结果作为最终的平滑度取值。具体数据开列在表 2 和表 3 中。

最终在构造样条曲线的时候使用的平滑度为：$s = c * s_0$ (11)

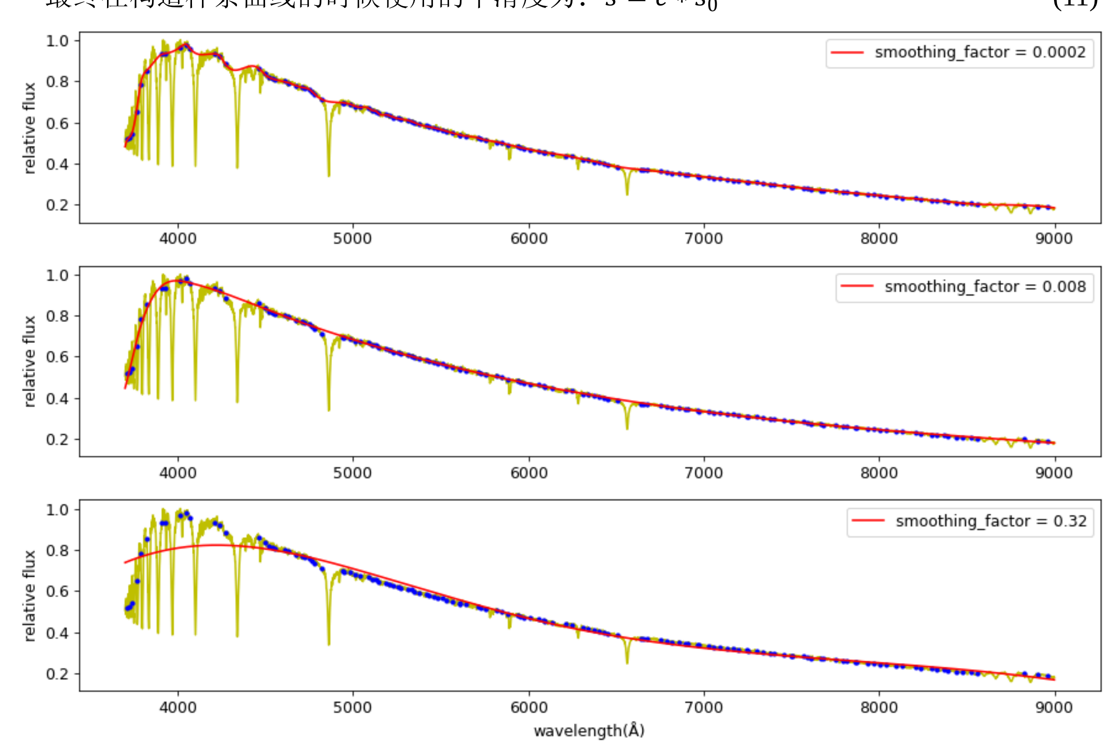


图 8 平滑度对拟合曲线的影响

Fig 8 The influence of smoothing factor on fitting curve

表 2 根据参考点集中流量最大值点所在波长位置选择平滑度

Tab. 2 Selection of smoothing factor according to the wavelength position of the maximum flux point in the reference point set

wave(A)	3700~4760	4760~6880	6880~7940	7940~9000
s_0	0.01	0.02	0.03	0.05

表 3 根据参考点集中数据点数 *n* 选择平滑度比例系数 *c*

Tab. 3 Selection of proportionality coefficient *c* of smoothing factor according to the number of data points *n* in the reference point set

n	1~29	30~59	60~89	90~119	120~149	150~180
c	0.2	0.35	0.5	0.65	0.8	1

chinaXiv:201903.00229v1

2　实验结果

2.1 不同情况下的处理结果

为检验所提出的方法的处理效果，从 LAMOST 中随机抽取了不同类型光谱 10000 条进行实验。

从中选出不同温度的光谱进行拟合，如图 9 所示。图中左侧蓝色实线为原始光谱，红色实线为拟合曲线；右侧绿色实线为对应的归一化谱。由图可见本文所提出的方法在不同温度下，均做到了较好的拟合。

如图 10 所示，左侧桃红色实线是原始光谱，绿色实线为样条函数拟合曲线；右侧靛蓝色实线为对应的归一化谱。选取不同的波长覆盖范围以后，本方法能够较好的工作，这使得本方法能够自动适用于需要局部归一化的场合，不需人工干预。

如图 11 所示，左侧是原始光谱及拟合曲线，在不同信噪比情况下，本方法拟合效果受影响不大。与文献[9]中专为低质量光谱开发的归一化方法处理效果非常接近。

如图 12 所示，在光谱中含有发射线及宇宙线的情况下，本方法仍能有效识别并进行准确的拟合。

2.2 误差分析和适用范围

由于仪器原因，LAMOST 光谱中常见的仪器形变模式有 3 种，如图 8 左侧子图所示。为了测试本文提出的方法是否能够有效地处理这些仪器导致的光谱形变，在有效温度 4000~8000K 的范围内随机抽取 9 条光谱，每条光谱归一化后的结果都乘以这 3 种形变曲线，并再次归一化。图 8 右侧子图所示的直方图中显示的是两次归一化结果残差的分布，可以看出本文算法的误差的平均值在 10^{-3} 量级，误差弥散在 10^{-2} 量级。

经实验分析，本文所提出的算法适用范围如下：5<SNR<600，波长覆盖范围 3700~9000Å，有效温度范围 3000~50000K。其中，信噪比在低于 5 的时候算法处理结果变得不稳定，总体精度有随着信噪比下降而降低的趋势。有效温度小于 3000K 的光谱处理结果同样不稳定，大量光谱总结果变得不可信。除此之外，信噪比高于 600，波长范围小于 3700Å 或大于 9000Å，有效温度高于 50000K 的光谱未做实验。并且，本算法适用于中低色散的光谱，高色散光谱需要对参数进行微调。

需要指出，对于本文算法涉及的参数：w, t,r_0,s_0,c.这些参数值的选择是为了适应仪器特性，对于其他来源的光谱，方法仍然可用但参数初值需要略微调整。

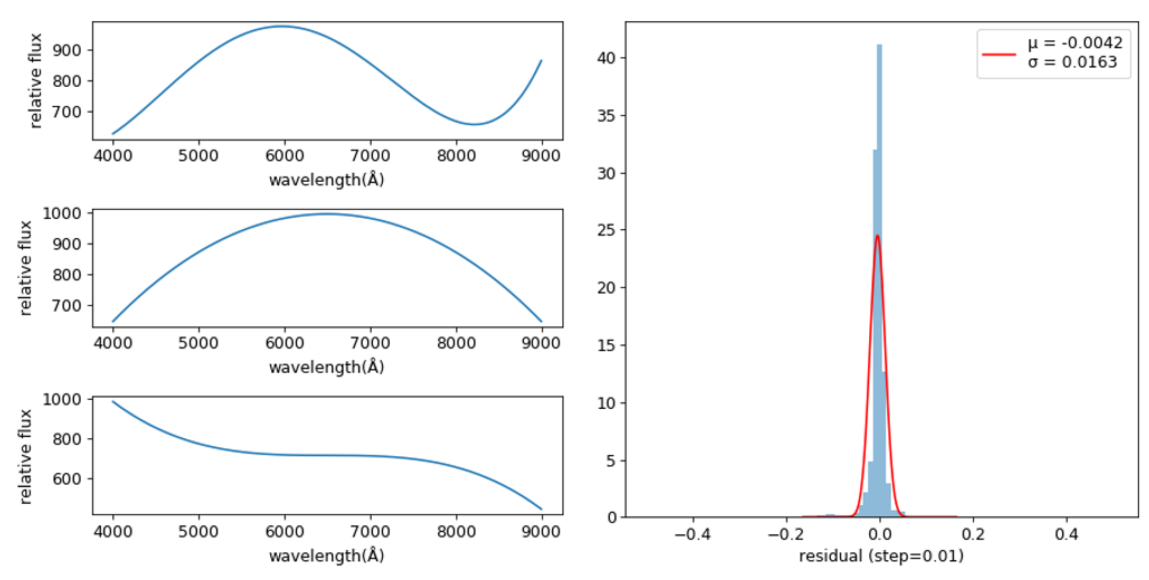


图 8 LAMOST 仪器的形变模式及归一化误差分析
Fig. 8Deformation mode in LAMOST and error analysis of normalization

ChinaXiv合作期刊

3 结论与展望

归一化是光谱数据处理中的重要环节，这一步骤处理结果的准确度将直接影响后续数据分析结果的正确性以及精确程度。为了获得更加准确的连续谱，在国内外研究的基础上提出了基于固定窗口划分的样条函数拟合方法，通过大量实验修正其中某些参数的经验取值。随后，使用更多的数据进行验证，结果表明本方法在拓宽适用范围的前提下实现了自动处理，与其他方法相比具有更好的普适性，在需要大规模自动化处理的场合有独特的优越性。

恒星光谱的归一化方法有很多种，在不同的场合下各有优劣。本文算法致力于实现LAMOST 海量光谱实时处理的自动化进行及普适性拓展。涉及到的参数值都是在大量实验的基础上，由经验总结出来的。另外，在开发算法的过程中，有一个重要问题未得到很好的解决。即缺乏一个严格的最优解评价标准来衡量归一化结果的好坏，使得算法仅能对一些明显的错误，如连续谱出现负值，做出基本的容错处理。如果能够研究出可量化并且方便计算机实现的归一化结果评价标准，则可以对处理结果进行评价，通过参数自动修改适应来再次处理直至获得最优解。

随着天文数据的急剧增长，我们还需要更加高效和智能的归一化方法及数据处理程序。下一步将进行低温星、特殊星如碳星等的光谱归一化研究。

chinaXiv:201903.00229v1

ChinaXiv合作期刊

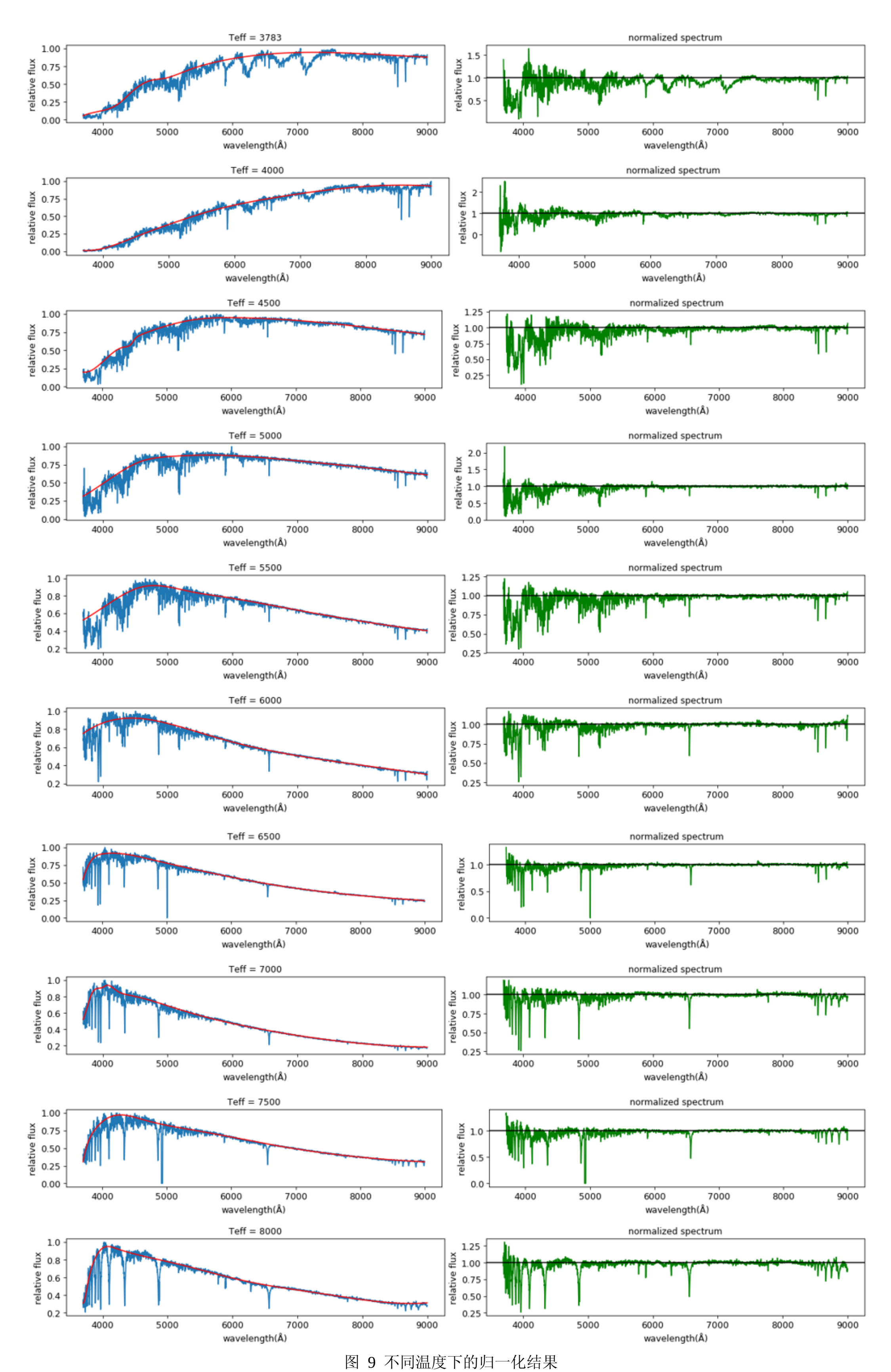


图 9 不同温度下的归一化结果
Fig. 9 Normalization results at different temperatures

chinaXiv:201903.00229v1

chinaXiv:201903.00229v1

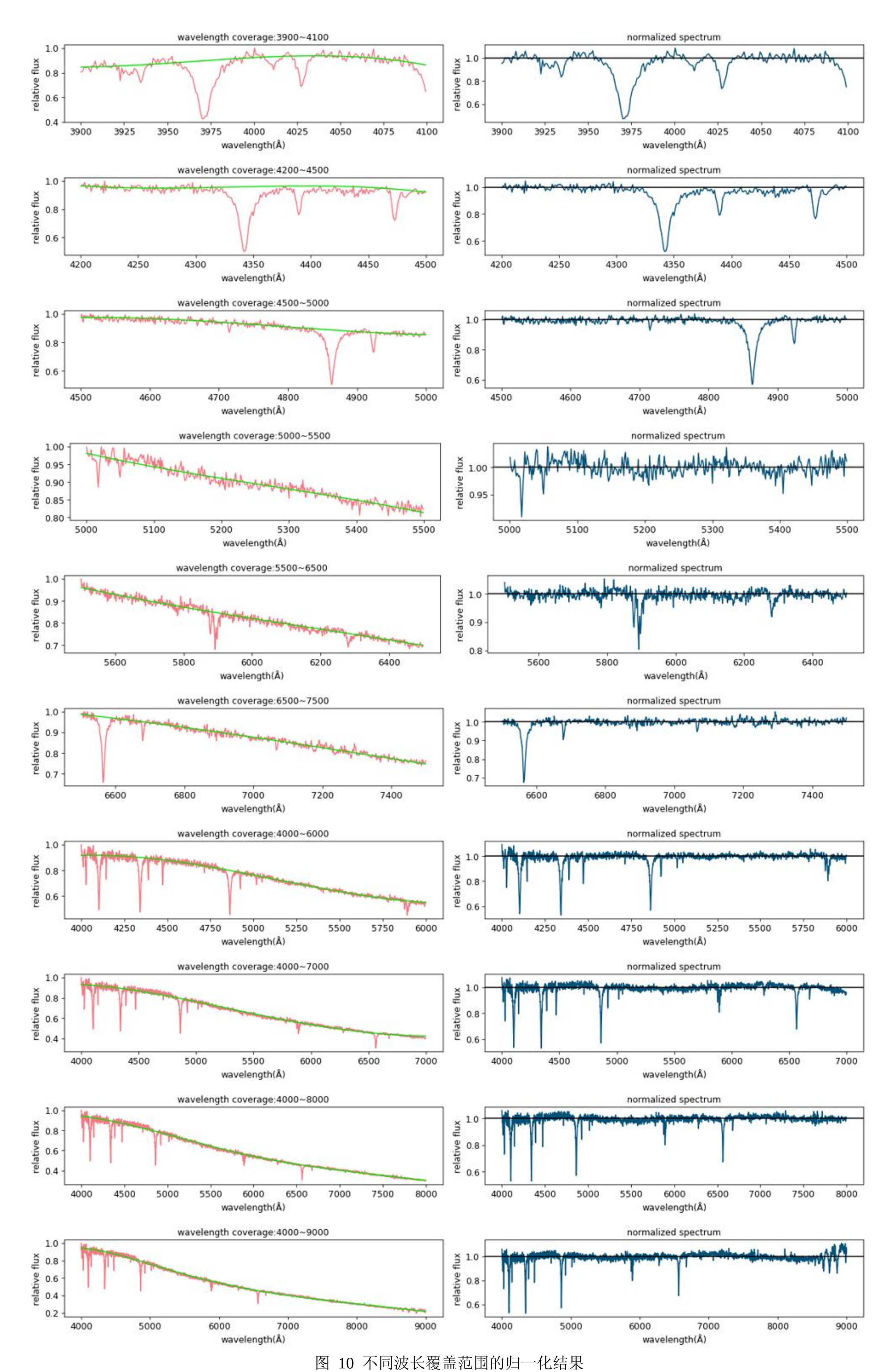


图 10 不同波长覆盖范围的归一化结果

Fig 10 Normalization results of different wavelength coverage range

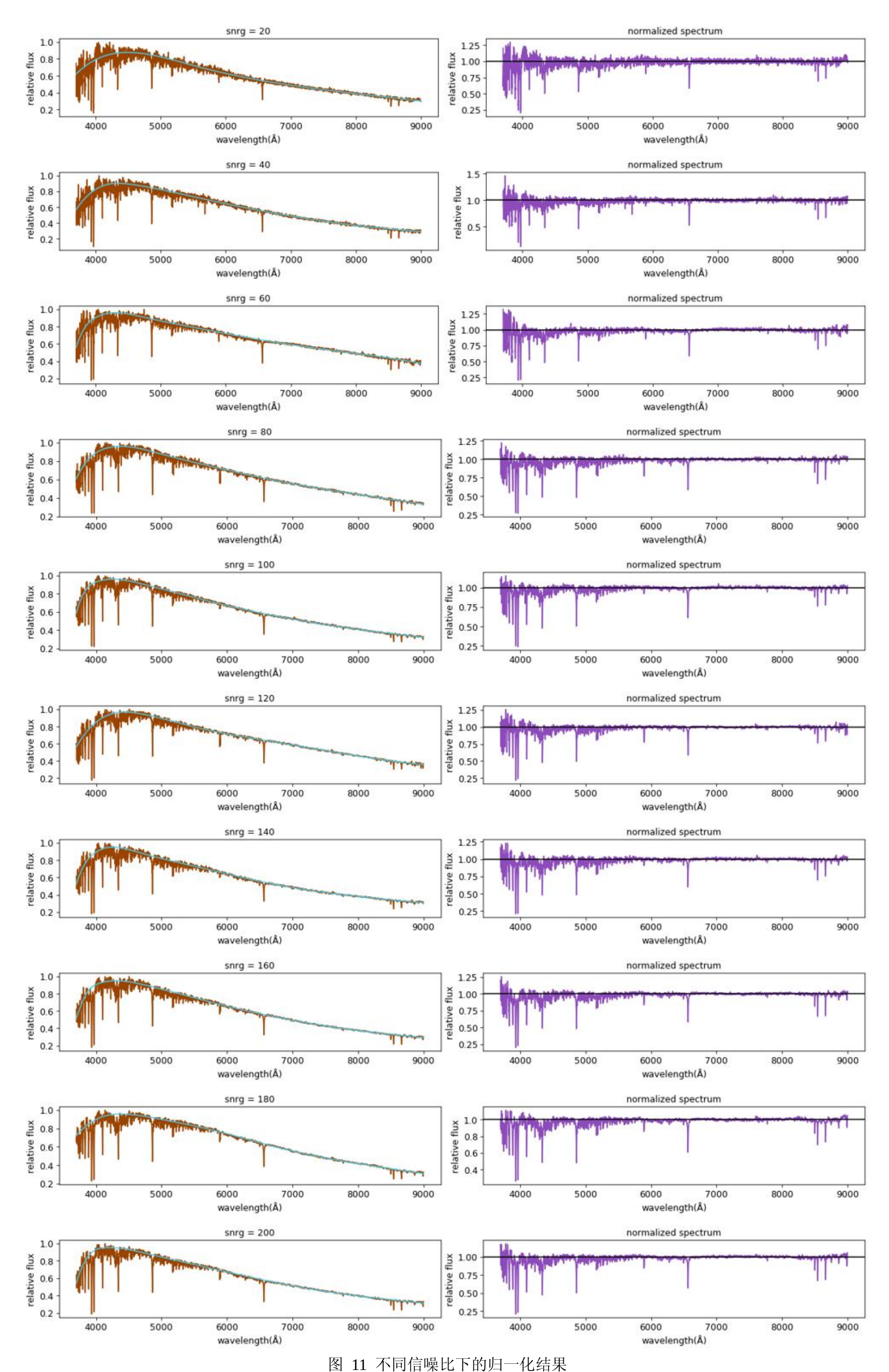


图 11 不同信噪比下的归一化结果
Fig 11 Normalization results under different SNR

chinaXiv:201903.00229v1

ChinaXiv合作期刊

chinaXiv:201903.00229v1

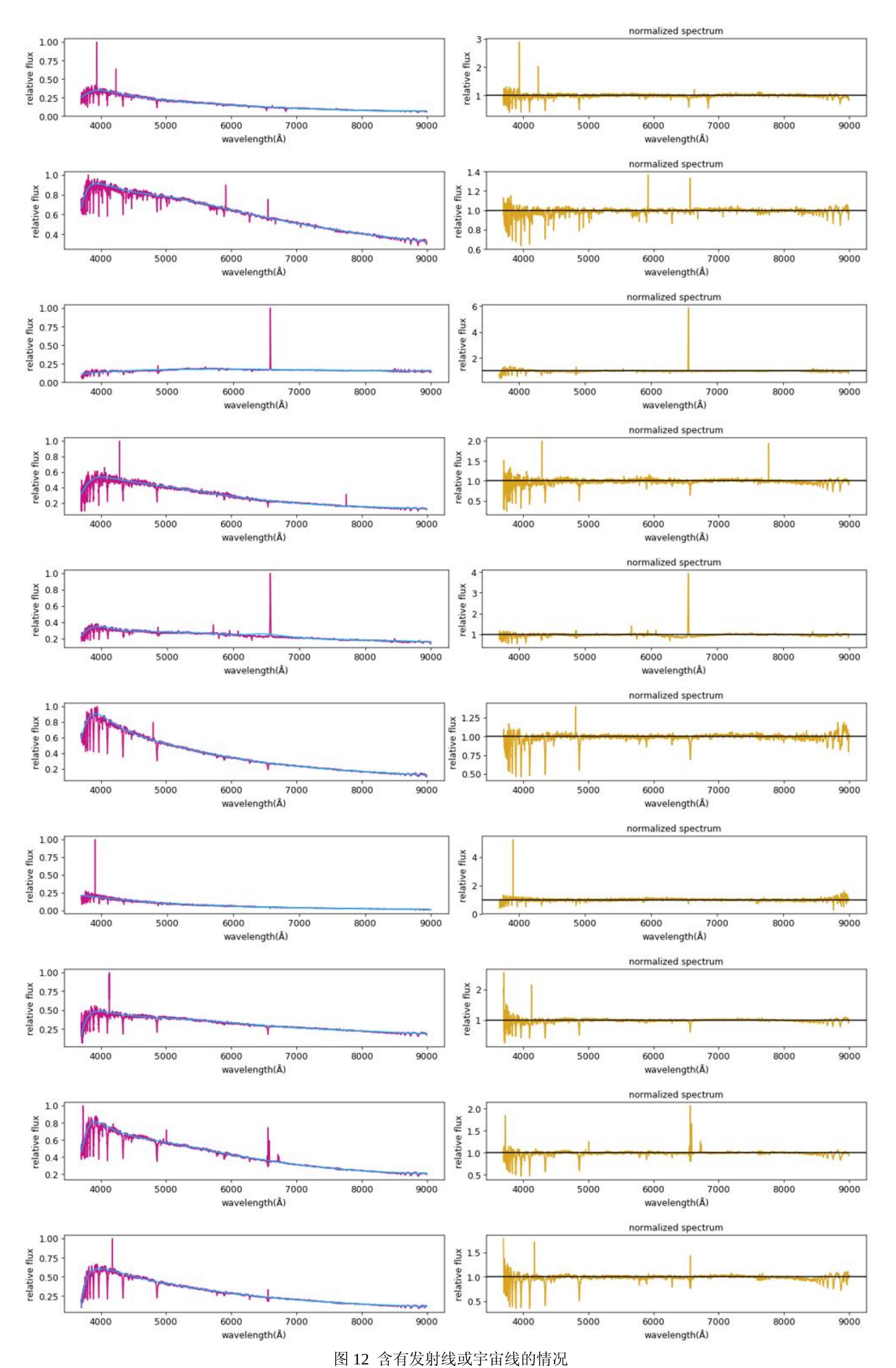


图 12 含有发射线或宇宙线的情况
Fig. 12 Circumstances containing emission lines or cosmic rays

ChinaXiv合作期刊

参考文献:

［1］LAMOST celestial spectral survey [J].Phsics(赵永恒. LAMOST天体光谱巡天[J]. 物理), 2015,44(04):205-212.
［2］ Automatic Normalization and Equivalent-Width Measurement of High-Resolution Stellar Spectra[J].Chinese Journal of Astronomy and Astrophysics,2006(06):689-696.
［3］Starck J L, Siebenmorgen R, Gredel R. The Astrophysical Journal, 1997, 482:1011.
［4］A Novel Method for Continuum Normalization of Astronomical Spectrum Signals(赵瑞珍,罗阿理.天体光谱信号的连续谱归一化新方法[J].光谱学与光谱分析),2006(03):587-590.
［5］ Lee Y S, Beers T C, Sivarani T, et al. The SEGUE stellar parameter pipeline. I. Description and comparison of individual methods[J]. The Astronomical Journal, 2008, 136(5): 2022.
［6］曾谨言. 量子力学: 卷 I[M]. 科学出版社, 2013:3-4
［7］The automatic detection of the continuum problem in the stellar spectra based on distance metric(于敬敬,潘景昌,孟凡龙,韦鹏.基于距离度量的LAMOST光谱中连续谱异常的自动检测[J].光谱学与光谱分析),2017,37(07):2246-2249.
［8］Schumaker L. Spline functions: basic theory[M]. Cambridge University Press, 2007.
［9］A method to fit low-quality stellar spectrum(吴明磊,潘景昌,衣振萍,韦鹏.恒星低质量光谱的连续谱拟合方法[J].光谱学与光谱分析),2018,38(03):963-967.

chinaXiv:201903.00229v1

ChinaXiv合作期刊

Automatic normalization method of stellar spectrum based on spline function

Luo Feng[1,2], Liu Chao[1,2], Zhao Yongheng[1,2].

(1,National Astronomical Observatory of China,Beijing 100101;2,University of Chinese Academy of Sciences,Beijing 100101)

Abstract: The observed spectrum of stars is generally composed of continuous spectrum, spectral line and noise.The continuous spectrum is the smooth continuous spectrum of radiation flux varying with wavelength caused by blackbody radiation.Spectral classification and stellar physical parameter estimation depend on the accurate extraction of continuous spectrum and spectral line information.Therefore, the main work of spectral data processing is to fit continuous spectrum and extract spectral line features by normalization.At present, the methods of continuous spectrum fitting mainly include polynomial fitting, median filtering and wavelet filtering.Existing methods have limitations to varying degrees in the case of low SNR, interference of cosmic ray signals and existence of emission lines, which are mainly reflected in robustness and accuracy.For the time being, there is no automated method applies to the normalization of the 10^7spectra from LAMOST.In the period of avalanche of astronomical data, it is very urgent to research and develop a spectral normalization algorithm of stars with better universality and automatic processing that can be applied to a wider range of temperature, SNR and wavelength coverage.On the basis of careful analysis of different types of spectra, a continuous spectrum fitting method based on fixed window dividing is proposed.This method can filter and extract the data points in the spectrum which can reflect the characteristics of continuous spectrum, and produce more accurate continuous spectrum by fine controlling the smoothness of spline function.Experiments were carried out using spectra of different spectral types, temperature ranges and wavelength coverage ranges in LAMOST, and the results showed that the proposed algorithm had good accuracyanduniversality.

Key words: Continuous spectrum normalization; LAMOST; stellar spectrum; Spline function

chinaXiv:201903.00229v1